COOLEY LLP
BOBBY GHAJAR (198719)
(bghajar@cooley.com)
COLETTE GHAZARIAN (322235)
(cghazarian@cooley.com)
1333 2nd Street, Suite 400
Santa Monica, California 90401
Telephone: (310) 883-6400

MARK WEINSTEIN (193043)
(mweinstein@cooley.com)
KATHLEEN HARTNETT (314267)
(khartnett@cooley.com)
JUDD LAUTER (290945)
(jlauter@cooley.com)
ELIZABETH L. STAMESHKIN (260865)
(lstameshkin@cooley.com)
3175 Hanover Street
Palo Alto, CA 94304-1130
Telephone: (650) 843-5000

CLEARY GOTTLIEB STEEN & HAMILTON LLP
ANGELA L. DUNNING (212047)
(adunning@cgsh.com)
1841 Page Mill Road, Suite 250
Palo Alto, CA 94304
Telephone: (650) 815-4131

PAUL, WEISS, RIFKIN, WHARTON & GARRISON LLP
KANNON K. SHANMUGAM (*pro hac vice*)
(kshanmugam@paulweiss.com)
2001 K Street, NW
Washington, DC 20006
Telephone: (202) 223-7300

Counsel for Defendant Meta Platforms, Inc.

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

SAN FRANCISCO DIVISION

RICHARD KADREY, *et al.*, Individual and Representative Plaintiffs, v.	Case No. 3:23-cv-03417-VC-TSH **DECLARATION OF JOELLE PINEAU IN SUPPORT OF META'S MOTION FOR PARTIAL SUMMARY**

DECLARATION OF JOELLE PINEAU
CASE NO. 3:23-CV-03417-VC-TSH

META PLATFORMS, INC., a Delaware corporation;	**JUDGMENT**
Defendant.	

I, Joelle Pineau, Ph.D., declare:

1.   I am over the age of 18 and am competent to make this declaration. I am currently a Vice President of AI Research in the Fundamental AI Research ("FAIR") organization of Meta Platforms, Inc. ("Meta"). I have been employed by Meta since May 2017. I have personal knowledge of the facts contained in this declaration in support of Defendant Meta Platform Inc.'s Motion for Partial Summary Judgment. I declare that the following is true to the best of my knowledge, information, and belief, and that if called upon to testify, I could and would testify to the following.

Professional Background

2.   I obtained my Bachelor's degree (BASc) in Engineering from the University of Waterloo, and my Masters (MASc) and Ph.D. from Carnegie Mellon University in Robotics.

3.   My job responsibilities at Meta include leading the FAIR organization. I joined Meta in 2017 as a Research Manager for AI Research. I have held my current role of Vice President of AI Research at Meta for approximately three years.

4.   I am also currently employed by McGill University as a Professor in Computer Science. I have been employed by McGill University since 2004. My academic research focuses on developing new models and algorithms for planning and learning in complex partially observable domains. I am a past President of the International Machine Learning Society, the inaugural Reproducibility Chair of the NeurIPS conference, and I am the creator of the ML Reproducibility checklist and the ML Reproducibility challenge. I am a recipient of NSERC's

E.W.R. Steacie Memorial Fellowship (2018), the Governor General's Innovation Awards (2019), a CIFAR Canada AI (CCAI) chair-holder, a Fellow of the Association for the Advancement of Artificial Intelligence (AAAI), and a Fellow of the Royal Society of Canada (RSC).
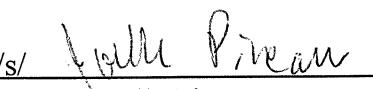
Use of AI Training Datasets

5.   Throughout my time with FAIR, Meta has downloaded and used a number of publicly-available text datasets to train its large language models (LLMs). These datasets are used in the research and development of LLMs, including training LLMs that are released or made available to the public (for example to the open-source community). But before an LLM that was trained using a particular dataset is made available to the public, Meta will typically use that dataset internally to perform experiments and analyses to assess the likely impact of that dataset on performance of the LLM. This often involves training an internal research LLM and determining how its performance would be affected (either positively or negatively) if a particular dataset were added to the data mix used for training. This type of research is broadly conducted within the AI research community and is, thus, not limited to Meta. Within Meta, this research is often performed by different teams and over different periods of time, for example, by teams who are studying different aspects of model behavior or who are evaluating use of a particular dataset for a particular model. This often results in particular datasets being separately acquired and separately used by different teams within Meta over time to facilitate the research.

6.   The process of acquiring and internally using datasets is essential to AI research and development (including LLM research) in order to ascertain what types of data are most likely to improve model performance and/or efficiency and which data may be redundant or unhelpful. This analysis not only contributes to a better understanding of the AI training process and the impact of certain types of data on model performance, but is often used to inform internal development or business decisions about the data or datasets that should be acquired and/or used

1   to train Meta's AI models. For example, in 2022, researchers at FAIR downloaded data from a

2   dataset known as Library Genesis (LibGen) and, by late 2022, used it to perform internal

     experiments to ascertain the impact using such data would have on performance of AI models, as

3   measured using various industry standard benchmarks. I am not aware of Meta having ever used

     LibGen for purposes other than the development and training of Meta's AI models, which

4   necessarily include these critical experiments and research.

5   I declare under penalty of perjury that the foregoing is true and correct. Executed on this

     17 day of April, at _Montreal, Canada_ .

6

7   /s/ _Joelle Pineau_
                                                 Joelle Pineau

8

9

10

11

12

13

14

15

16

17   4   DECLARATION OF JOELLE PINEAU
                                                               CASE NO. 3:23-CV-03417-VC-TSH